Systems for Generating and Analyzing Stimulus-Response Output Signal Matrices

Inventors: Jasper Rine and Mathew Ashby
Assignee: Regents of the University of California
Patent Attorney: Richard Aron Osman, Ph.D., Reg. No. #36,627

INTRODUCTION

SUB A1

Field of the Invention

The field of the invention is the generation and analysis of stimulus-response signal profiles adapted to computer-based artificial intelligence systems such as neural networks and expert systems and their use as models for systemic responses.

Background

Artificial intelligence (AI) systems can integrate data accumulation, recognition and storage functions with higher order analysis and decision protocols. AI systems such as expert systems and neural networks find wide application in qualitative analysis. Expert systems typically generate an individual data structure which is analyzed according to a knowledge base working in conjunction with a resident database; see, e.g. Holloway et al. (1993) US PATENT NO 5,253,164 which was subject to recent judicial review, GMIS Inc. 34 USPQ2d 1389 (1995). "MYCIN", another example, is a computer protocol using individual clinical evaluations to generate a personal data structure which is analyzed according to a knowledge base to predict or diagnose myocardial infarction and to determine hospital admissibility (Goldman et al. (1988) New England Journal of Medicine 318, 797-803).

Neural network systems are networks of interconnected processing elements, each of which can have multiple input signals, but generates only one output signal. A neural network is trained by inputting training set of signals and correlating responses. The trained network is then used to analyze novel signals. For example, neural networks have been used extensively in optical character and speech recognition applications (e.g. Colley et al. (1993) US Patent No. 5,251,268).

The analysis of complex systems such as biological organisms are particularly well-suited to AI systems. Otherwise intractable complex stimulus-response patterns can be effectively analyzed using deduction protocols applied through AI systems. Pharmaceutical development for example, requires large-scale studies of systemic responses to modifications of the structure, form or administration of a drug. Presently, such systemic information is usually provided by live animal models which are costly and provide limited, relatively uninformative output signals (death, weight loss, etc.) and often mask the myriad biochemical pathway repressions and activations which underlie the measured organismal response. A number of in vitro or cell culture-based methods have been described for identifying compounds with a particular biological effect through the activation of a linked reporter (e.g. Gadski et al. (1992) EP 92304902.7 describes methods for substances which regulate the synthesis of an apolipoprotein; Evans et al. (1991) US Patent No. 4,981,784 describes methods for identifying ligand for a receptor and Farr et al. (1994) WO 94/17208 describes methods and kits utilizing stress promoters to determine toxicity of a compound).

The present invention combines these approaches to provide an in vitro or cell culture-based analysis of systemic response patterns. In particular, the invention involves sophisticated methods for generating and analyzing highly informative stimulus - systemic repression and activation response patterns.

SUMMARY OF THE INVENTION

The invention provides systems and methods for generating an output signal matrix database and for analyzing an output signal matrix by comparison to an output signal matrix database for correlating candidate stimuli and responses.

Generating an output signal matrix database according to the invention involves: (I) constructing a stimulated physical matrix; (ii) detecting a physical signal at each unit of the physical matrix; (iii) transducing each physical signal to generate a corresponding electrical output signal; (iv) storing each output signal in an output signal matrix data structure associating each output signal with the X and Y coordinates of the corresponding physical matrix unit and the stimulus; and (v) repeating steps (I) - (iv) to iteratively store output signal matrix data structures for a plurality of stimuli to form an output signal matrix database indexing output signal matrix data structures by stimuli.

Analyzing an output signal matrix by comparison to an output signal matrix database according to the present invention involves: (a) constructing a stimulated physical matrix; (b) detecting a physical signal at each unit of the physical matrix; (c) transducing each physical signal to generate a corresponding electrical output signal; (d) storing each output signal in an output signal matrix data structure associating each output signal with the X and Y coordinates of the corresponding physical matrix unit and the stimulus; and (e) comparing the output signal matrix data structure of step (d) with an output signal matrix database produced by the foregoing method of generating an output signal matrix database.

The stimulated physical matrices comprise an ordered array of units having X and Y coordinates. Each unit confines (1) either a different responder of a living thing or a probe corresponding to such a different responder and, (2) an identifier for the responder or probe. The living thing is provided a stimulus capable of repressing the responders of a plurality of the units and the identifier provides a physical signal corresponding to the repression of such different responder. The arrays may comprise the organism's entire repertoire of responders which may be genes, gene regulatory elements, gene transcripts or gene translates, or a predetermined functional class or subset of the organism's entire repertoire. In a preferred embodiment, the array comprises a sufficient ensemble of responders to deduce the action of a stimulus regardless of its mechanism of action.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 is an overview of a system of a matrix analysis system.

Figure 2 is a flow chart representing the steps performed in generating an output signal matrix database.

Figure 3 is diagram representing a stimulated physical matrix.

Figure 4 is a flow chart representing the steps performed in generating an output signal matrix data structure and comparing said data structure with an output signal matrix database.

Figure 5 is a schematic of an expert system of design which may be used for analyzing an output signal profile.

Figure 6 is a flow chart representing the steps performed in generating a gene reporter matrix output response profile for an unknown stimulus, regulation tables, basal reference response profiles, known chemical response profiles, and known genetic response profiles.

Figure 7 is a flow chart representing the steps performed in analyzing a gene reporter matrix output response profile for an unknown stimulus.

DETAILED DESCRIPTION OF THE INVENTION

The present invention relates to methods for generating and analyzing biological stimulus-response patterns using deduction protocols applied through AI systems such as expert systems and neural networks. The systems may be used, for example, as in vitro or cell culture substitutes for live animal studies of drug efficacy. Figure 1 shows an overview of a system according to the invention. The system 100 includes a central processing unit 110, computer memory 122, user interface 114, system communication bus 112, and the physical matrix output signal transduction subsystem 116. The subsystem 116 includes a stimulated physical matrix 120 and a physical matrix output signal detector and signal transducer 118. The computer memory 122 contains collected output signal matrix data structures in the form of an output signal matrix database 124, a comparison function 126, and often, a knowledge base 128.

Figure 2 provides a schematic representation of the steps performed in generating an output signal matrix database indexing N output signal matrix data structures by corresponding stimuli. The first step 205 involves assigning N an integer value of one for the generation of the first output signal matrix data structure corresponding to the first stimulus.

The second step 210 of the method shown in Figure 2 involves constructing a stimulated physical matrix. Figure 3 provides a schematic representation of a stimulated physical matrix. The stimulated physical matrix 310 comprises an ordered array of units having X and Y coordinates. While Figure 3 depicts four illustrative units 312, in practice, the matrix will typically have about a hundred or more units. The units are generally a region of a solid substrate such as a two-dimensional portion of the surface of a silicon-based wafer, a well of a microtiter plate, etc. Each unit confines either a different responder 314 of a living thing or a probe 316 corresponding to such a different responder and, an identifier 318 for the responder or probe. Generally, all the units of a given matrix will employ a responder or all will have a probe. Further, for most convenient detection and data processing, all the units of a given matrix generally use the same identifier.

The living thing (or, organism) is provided a stimulus capable of repressing the responders 314 of a plurality of the units 312 and the identifier 318 provides a physical signal corresponding to the repression of such different responder 314. Responses, usually cellular, to a wide variety of stimuli may be monitored. Examples of stimuli include candidate pharmacological agents, suspected pathogenic agents, transfected nucleic acids, radiative energy, etc. The stimulus induces the repression of a plurality of the responders of the matrix, relative to their pre-stimulus state of induction, as measured by the pre-stimulus output signal at the corresponding unit. Typically, the stimulus provides a complex response pattern of repression, silence and induction across the matrix. The response profile reflects the cells' transcriptional adjustments to maintain homeostasis in the presence of the drug. Hence, while a wide variety of stimuli may be evaluated. it is important to adjust the incubation conditions (e.g. stimulus intensity, exposure time, etc.) to preclude cellular stress, and hence insure the measurements of pharmaceutically relevant response profiles. The arrays may comprise the organism's entire repertoire of responders which may be genes, gene regulatory elements, gene transcripts or gene translates (proteins), or a predetermined functional class or subset of the organism's entire repertoire. By incorporating at least 0.5%, preferably at least 5%, more preferably at least half, most preferably essentially all the responders (e.g. gene regulatory regions) of the organism, an in vitro or cell culture model of the organism (e.g. animal) may be obtained. In a preferred embodiments, the array comprises a sufficient ensemble of responders so as to model the systemic response of the organism and to deduce the action of a stimulus regardless of its mechanism of action.

The nature of the linkage 320 between the responder and identifier will vary with the application of the matrix. As examples: each unit of a matrix reporting on gene expression might confine a cell having a construct of a reporter gene operatively joined to a different transcriptional promoter; alternatively, each unit of a matrix reporting on gene expression might confine a different oligonucleotide probe capable of hybridizing with a corresponding different reporter transcript; each unit of a matrix reporting on DNA-protein interaction might confine a cell having a first construct of a reporter gene operatively joined to a targeted transcription factor binding site and a second hybrid construct encoding a transcription activation domain fused to a different structural gene (a one-dimensional one-hybrid system matrix); each unit of a matrix reporting on protein-protein interactions might confine a cell having a first construct of a reporter gene operatively joined to a targeted transcription factor binding site, a second hybrid construct encoding a transcription activation domain fused to a different constitutionally expressed gene and a third construct encoding a DNA-binding domain fused to yet a different constitutionally expressed gene (a two-dimensional two-hybrid system matrix).

The third step 212 of the method shown in Figure 2 involves detecting a physical signal at each unit of the physical matrix. The physical signal will depend on the nature of the responder and/or identifier. Typically, the signal is a change in one or more electromagnetic properties, particularly optical properties at the unit. As examples, a reporter gene may encode an enzyme which catalyzes a reaction at the unit which alters light absorption properties at the unit, radiolabeled or fluorescent tag-labeled nucleotides can be incorporated into nascent transcripts which are then identified when bound to oligonucleotide probes, etc. Electronic detectors for optical, radiative, etc. signals are commercially available. For example, Figure 1 shows an automated, multi-well colorimetric detector 118, similar to automated ELISA readers, reading a resident microtiter plate-type physical matrix 120. Generally, the method involves detecting a signal of the reporter at each unit of the matrix in response to the candidate stimulus or stimuli at a first intensity and again at a second intensity. Frequently, one of the intensities will be zero or at least below the threshold necessary for an effect on any of the cells. For example, signal may be detected before and after the cells of the matrix are incubated with a candidate pharmacological agent for a time and under conditions sufficient for the cells to respond to the presence of the agent. The signal may also be monitored as a function of other variables such as stimulus intensity or duration, time (for dynamic response analyses), etc.

The fourth step 214 of the method shown in Figure 2 involves transducing the physical signal at each unit to generate a corresponding electrical output signal. Generally, the signal transduction is a linear electronic conversion to a digital signal. Electronic converters capable of linear signal transduction are commercially available. The physical matrix output signal detector and signal transducer 118 shown in Figure 1 houses this transducing function.

The fifth step 216 of the method shown in Figure 2 involves storing in memory each output signal in an output signal matrix data structure associating each output signal with the X and Y coordinates of the corresponding physical matrix unit and the stimulus. As depicted in Figure 1, the data structure may be stored in the form of an output signal matrix database 124 in the memory 122 of a computer 110.

After the fifth step of the method shown in Figure 2, the routine advances to a decision block 218. If all the requisite data structures have not been completed, the routine advances to step 220 where n is incremented and the data structure generation steps 210-216 are repeated for the n+1th stimulated physical matrix. If, after the fifth step of the method shown in Figure 2, all the requisite data structures have been completed, the routine advances to step 222. Step 222 involves forming an output signal matrix database indexing output signal matrix data structures by stimuli. This compilation step usually involves digitizing an analog signal corresponding to the reporter signal at each unit, pairing each digitized signal with a matrix unit identifier and storing said pairs as an output signal database in the memory of a computer.

Figure 4 provides a schematic representation of the steps performed in generating an output signal matrix data structure and comparing said data structure to an output signal matrix database indexing N output signal matrix data structures by corresponding stimuli.

The first four steps 410, 412, 414 and 416 are as described for steps 210, 212, 214 and 216 of Figure 2. In particular, step 410 of the method involves constructing a stimulated physical matrix; the second step 412 involves detecting a physical signal at each unit of the physical matrix; the third step 414 involves transducing each physical signal to generate a corresponding electrical output signal; and, the fourth step 416 of the method involves storing in memory each output signal in an output signal matrix data structure. The fifth step 418 of the method involves comparing the output signal matrix data structure of step (d) with an output signal matrix database produced by the foregoing method of generating an output signal matrix database. The comparison step is generally performed by computer system employing AI technology. For example, the system 100 of Figure 1 shows a central processing unit and a user interface 114 working in conjunction with a memory 122 which stores the output signal matrix database 124 and the comparison function 126. Here, the comparison function 126 provides the AI technology for comparing, for example, an unknown output signal matrix with a database 124 to deduce the mechanism of action and characteristics of the responsible stimulus. For example, an expert system using an input knowledge base of comparison rules or a neural network trained on a population of known stimulus-response pairs may be used.

In particular, Figure 5 provides a schematic representation of an expert system 500 which may be used to compare an output signal matrix data structure of step with an output signal matrix database. Such a system may be implemented in hardware or software. The knowledge base 510 comprises an output signal matrix database and a series of comparison rules. The system interface 514 permits the input of library data structures for the database and query data structures or query stimuli. The inference engine 512 is a computer program that processes data structures for comparison against the resident knowledge base database according to the knowledge base rules to generate correlates and qualitative and/or quantitative deduction analyses. Such analyses are conveniently output as a prioritized set of matches, each match including an identifier and a relatedness score as in Basic Local Alignment String Transformer (BLAST) search reports, Altschul et al. (1990) Basic Local Alignment Search Tool, J Mol Biol 215, 403-410.

A particular embodiment of the invention is a substantially comprehensive gene reporter matrix, or genome reporter matrix. Such a substantially comprehensive matrix includes at least a majority of the organism's genes and, in a preferred embodiment, essentially all different genes of the target organism. Because yeast, such as *Saccharomyces cerevisiae*, is a bona fide eukaryote, there is substantial conservation of biochemical function between yeast and human cells in most pathways, from the sterol biosynthetic pathway to the Ras oncogene. Indeed, the absence of many effective antifungal compounds illustrates how difficult it has been to find therapeutic targets that would preferentially kill fungal but not human cells.

One example of a shared response pathway is sterol biosynthesis. In human cells, the drug Mevacor (lovastatin) inhibits HMG-CoA reductase, the key regulatory enzyme of the sterol biosynthetic pathway. As a result, the level of a particular regulatory sterol decreases, and the cells respond by increased transcription of the gene encoding the LDL receptor. In yeast, Mevacor also inhibits HMG-CoA reductase and lowers the level of a key regulatory sterol. Yeast cells respond in an analogous fashion to human cells. However, yeast do not have a gene for the LDL receptor. Instead, the same effect is measured by increased transcription of the ERG10 gene, which encodes acetoacetyl CoA thiolase, an enzyme also involved in sterol synthesis. Thus the regulatory response is conserved between yeast and humans, even though the identity of the responding gene is different.

The genome reporter matrix is exemplified below with a cell-based gene product reporter system; however, the other types of stimulated physical matrices described herein may be used. A genome reporter matrix comprising a comprehensive collection of reporter genes for about 6,000 genes of the yeast signals the regulatory response of a single gene to essentially anything that influences the function of the cell. For example, an increased production of the ERG10 reporter would signal the presence of an inhibitor of sterol synthesis. In practice, a yeast genome reporter matrix may be made with a collection of several thousand yeast strains each of which contains a single gene fusion of a yeast gene to a reporter gene. In one manifestation, the reporter fusions will be hybrid genes in which the lacZ gene is fused to a yeast gene, so that the activity of the yeast gene promoter directs synthesis of β-galactosidase. These fusion gene-containing strains are conveniently arrayed into microtiter plates in liquid culture and a permanent collection maintained at -80°C. Copies of this collection can be made and propagated by simple mechanics and may be automated with commercial robotics. The addition of an inhibitor to the entire collection of strains results in no change in the expression of the reporter in some strains, increased expression in some strains and decreased expression in others. By knowing the identity of the gene being reported in each strain, the genome-wide response is interpreted.

The strengths of this procedure are best illustrated by examples. Consider the difference between an in vitro assay for HMG-CoA reductase inhibitors as presently practiced by the pharmaceutical industry, and an assay for inhibitors of sterol biosynthesis as revealed by the ERG10 reporter. In the case of the former, information is obtained only for those rare compounds that happen to inhibit this one enzyme. In contrast, in the case of the ERG10 reporter, any compound that inhibits nearly any of the approximately 35 steps in the sterol biosynthetic pathway will induce the synthesis of the reporter. Thus, the reporter can detect a much broader range of targets than can the purified enzyme, in this case 35 times more than the in vitro assay.

Drugs often have side effects that are in part due to the lack of target specificity. However, the in vitro assay of HMG-CoA reductase provides no information on the specificity of a compound. In contrast, a genome reporter matrix reveals the spectrum of other genes in the genome also affected by the compound. In considering two different compounds both of which induce the ERG10 reporter, if one compound affects the expression of 5 other reporters and a second compound affects the expression of 50 other reporters, the first compound is, a priori, more likely to have fewer side effects. Because the identities of the reporters are known or determinable, information on other affected reporters is informative as to the nature of the side effect. A panel of reporters can be used to test derivatives of the lead compound to determine which of the derivatives have greater specificity than the first compound.

As another example, consider the case of a compound that does not affect the in vitro assay for HMG-CoA reductase nor induces the expression of the ERG10 reporter. In the traditional approach to drug discovery, a compound that does not inhibit the target being tested provides no useful information. However, a compound having any significant effect on a biological process generally has some consequence on gene expression. A genome reporter matrix can thus provide two different kinds of information for most compounds. In some cases, the identity of reporter genes affected by the inhibitor evidences to how the inhibitor functions. For example, a compound that induces a cAMP-dependent promoter in yeast may affect the activity of the Ras pathway. Even where the compound affects the expression of a set of genes that do not evidence the action of the compound, the matrix provides a comprehensive assessment of the action of the compound that can be stored in a database for later analyses. A library of such response profiles can be continuously investigated, much as the Spectral

Compendiums of chemistry are continually referenced in the chemical arts. For example, if the database reveals that compound X alters the expression of gene Y, and a paper is published reporting that the expression of gene Y is sensitive to, for example, the inositol phosphate signaling pathway, compound X is a candidate for modulating the inositol phosphate signaling pathway. In effect the genome reporter matrix is an informational translator that takes information on a gene directly to a compound that may already have been found to affect the expression of that gene. This tool should dramatically shorten the research and discovery phase of drug development, and effectively leverage the value of the publicly available research portfolios on all genes. In contrast to the existing approaches that require the continual development of different assays, the genome reporter matrix measures the effect of all inhibitor with the same assay.

A genome reporter matrix may consist of three classes of reporters: One hundred reporter gene fusions constructed by PCR technology to that subset of genes that are understood and whose expression is most revealing of physiological changes throughout the cell; an arrayed random library sufficiently complex to include a fusion to most yeast genes; and reporters to key genes in a collection of genetically sensitized strains containing mutations that sensitize that strain to effects on a certain pathway. For example, the first class of reporters is designed as perfect lacZ translational fusions in which the β-galactosidase coding sequence is fused to the initiation codon in the gene of interest. Such fusions are able to detect changes in the level of expression due either to changes in transcription or changes in translation. The genes detected by this method include the HO gene, to monitor effects on cell cycle function; the FUS1 gene, to detect changes in one of the map kinase cascades; ERG10 and HMG-CoA reductase, to report changes in the sterol biosynthetic pathway; GCN4, to report levels of charged amino acyl tRNAs; Ty1, to report levels of retroposon expression, HIS3 expression to report effects on the yeast JUN oncogene homolog; HMG2 to report inhibition of heme biosynthesis; genes that report the activity of the RAS pathway, other map kinase cascades, the level of phospholipid activity, etc.

In many cases, a drug of interest would work on protein targets whose impact on gene expression would not be known a priori. For example, taxol, a recent advance in potential breast cancer therapies, is thought to function by interfering with tubulin-based cytoskeletal elements. Even without knowledge of which genes in the genome reporter matrix are induced or repressed by inhibitors of the tubulin-based microtubule cytoskeleton, the genome reporter matrix can be used to determine this. Specifically, a dominant mutant form of tubulin is introduced into all the strains of the genome reporter matrix and the effect of the dominant mutant, which interferes with the microtubule cytoskeleton, evaluated for each reporter. This genetic assay informs us which genes would be affected by a drug that has a similar mechanism of action. In the case of taxol, the drug itself could be used to obtain the same information. However, the example demonstrates that even if taxol itself were not available, genetics can be used to predetermine what its response profile would be in the genome reporter matrix. Furthermore, it is not necessary to know the identity of any of the responding genes. Instead, the genetic control with the mutant tubulin sorts the genome into those genes that respond and those that do not. Hence, if drugs that disrupt the actin cytoskeleton were desired, dominant actin mutants introduced into the genome reporter matrix reveal what response profile to expect for such an agent.

Among the most important advances in drug development have been advances in combinatorial synthesis of chemical libraries. In conventional drug screening with purified enzyme targets, combinatorial chemistries can often help create new derivatives of a lead compound that will also inhibit the target enzyme but with some different and desirable property. However, conventional methods would fail to recognize a molecule having a substantially divergent specificity. The genome reporter matrix offers a simple solution to recognizing new specificities in combinatorial libraries. Specifically, pools of new compounds are tested as mixtures across the matrix. If the pool has any new activity not present in the original lead compound, new genes are affected among the reporters. The identity of that gene provides a guide to the target of the new compound. Furthermore, the matrix offers an added bonus that compensates for a common weakness in most chemical syntheses. Specifically, most syntheses produce the desired product in greatest abundance and a collection of other related products as contaminants due to side reactions in the synthesis. Traditionally the solution to contaminants is to purify away from them. However, the genome reporter matrix exploits the presence of these contaminants. Syntheses can be adjusted to make them less specific with a greater number of side reactions and more contaminants to determine whether anything in the total synthesis affects the expression of target genes of interest. If there is a component of the mixture with the desired activity on a particular reporter, that reporter can be used to assay purification of the desired component from the mixture. In effect, the reporter matrix allows a focused survey of the effect on single genes to compensate for the impurity of the mixture being tested.

Isoprenoids are a specially attractive class for the genome reporter matrix. In nature, isoprenoids are the champion signaling molecules. Isoprenoids are derivatives of the five carbon compound isoprene, which is made as an intermediate in sterol biosynthesis. Isoprenoids include many of the most famous fragrances, pigments, and other biologically active compounds, such as the antifungal sesquiterpenoids. There are roughly 10,000 characterized isoprene derivatives and many more potential ones. Because these compounds are used in nature to signal biological processes, they are likely to include some of the best membrane permeant molecules.

Isoprenes possess another characteristic that lends itself well to drug discovery through the genome reporter matrix. Pure isoprenoid compounds can be chemically treated to create a wide mixture of different compounds quickly and easily, due to the particular arrangement of double bonds in the hydrocarbon chains. In effect, isoprenoids can be mutagenized from one form into many different forms much as a wild-type gene can be mutagenized into many different mutants. For example, vitamin D used to fortify milk is produced by ultraviolet irradiation of the isoprene derivative known as ergosterol. New biologically active isoprenoids are generated and analyzed with a genome reporter matrix as follows. First a pure isoprenoid such as limonene is tested to determine its response profile across the matrix. Next, the isoprenoid (e.g. limonene) is chemically altered to create a mixture of different compounds. This mixture is then tested across the matrix. If any new responses are observed, then the mixture has new biologically active species. In addition the identity of the reporter genes provides information regarding what the new active species does, an activity to be used to monitor its purification, etc. This approach may be practiced in other different mutable chemical families in addition to isoprenoids.

Fungi are important pathogens on plants and animals and make a major impact on the production of many food crops and on animal, including human, health. In the development of antifungal compounds, one of the major goals has been to determine which targets are specific to the infecting fungus. The genome reporter matrix offers a new tool to solve this problem. Specifically, a reporter library is created from the targeted pathogen such as Cryptococcus, Candida, Aspergillus, Pneumocystis, etc. It is not necessary to know the identity of each reporter gene nor that the genome of the target species be sequenced. Compounds are tested for those that have an effect on the expression of any gene in the fungus. These positive compounds are then tested for those that have no or very little effect in Saccharomyces. By sequencing the reporter genes affected specifically in the target fungus and comparing the sequence with others in Genbank, one can identify biochemical pathways that are unique to the target species. Useful identified products include not only agents that kill the target fungus but also the identification of specific targets in the fungus for other pharmaceutical screening assays.

The identification of compounds that kill bacteria has been successfully pursued by the pharmaceutical industry for decades. It is rather simple to spot a compound that kills bacteria in a spot test on a petri plate. However, there is much complexity to bacterial physiology and ecology that could offer an edge to development of combination therapies for bacteria, even for compounds that do not actually kill the bacterial cell. Consider for example the bacteria that invade the urethra and persist there through the elaboration of surface attachments known as fimbrae. Antibiotics in the urine stream have limited access to the bacteria because the urine stream is short-lived and infrequent. However, if one could block the synthesis of the fimbrae to detach the bacteria, existing therapies would become more effective. Similarly, if the chemotaxis mechanism of bacteria were crippled, the ability of bacteria to establish an effective infection would, in some species, be compromised. A genome reporter matrix for a bacterial pathogen that contains reporters for the expression of genes involved in chemotaxis or fimbrae synthesis, as examples, identifies not only compounds that do kill the bacteria in a spot test, but also those that interfere with key steps in the biology of the pathogen. These compounds would be exceedingly difficult to discover by conventional means.

A genome reporter matrix based on human cells provides many important applications. For example, an interesting application is the development of antiviral compounds. When human cells are infected by a wide range of viruses, the cells respond in a complex way in which only a few of the components have been identified. For example, certain interferons are induced as is a double-stranded RNase. Both of these responses individually provides some measure of protection. A matrix that reports the induction of interferon genes and the double stranded RNase is able to detect compounds that could prophylactically protect cells before the arrival of the virus. Other protective effects may be induced in parallel. The incorporation of a panel of other reporter genes in the matrix is used to identify those compounds with the highest degree of specificity.

The methods involved in using a genome reporter matrix as presently described are essentially as described above. Specifically, the steps performed in generating and analyzing such a genome reporter matrix output signal database indexing N output signal matrix data structures by corresponding stimuli parallel the general method steps outlined in Figures 1 and 4. The first and second steps involve assigning n an integer value of one for the generation of the first output signal matrix data structure, and constructing a stimulated physical genome reporter matrix, respectively. In this embodiment, the stimulated physical matrix typically comprises a microtiter plate having 96 wells ordered in X and Y coordinates. Each well confines a cell or colony of cells having a construct of a reporter gene operatively joined to a different transcriptional promoter. The cells are provided a stimulus capable of repressing the promoters and thereby reducing the reporter expression in a plurality of the wells. The third step of the gene reporter matrix method involves detecting a physical signal resulting from reporter expression at each well. Generally, the reporter gene encodes an enzyme such as lacZ which provides a reaction product conveniently detected by spectroscopy. The fourth step involves linearly transducing the optical reporter signal at each well to generate a corresponding digital electrical output signal, and the fifth step involves storing each electrical output signal in computer memory as a gene reporter output signal matrix data structure associating each output signal with the coordinates of the corresponding microtiter plate well and the stimulus. After the fifth step, the routine advances to a decision block: if all the requisite data structures have not been completed, N is incremented and the data structure generation steps one to five are repeated for the N+1th stimulated genome reporter matrix. If, after the fifth step, all the requisite data structures have been completed, the routine advances to step of forming an output signal matrix database indexing output signal matrix data structures by stimuli. For analysis of an unknown stimulus (e.g. candidate drug), an AI system such as described previously is used to compare, usually a dilution series, of responses to the unknown with the database.

Figure 6 shows the steps performed in generating a gene reporter matrix output response profile for an unknown stimulus, regulation tables, basal reference response profiles, known chemical response profiles, and known genetic response profiles.

To generate a genome reporter matrix 610 a set of lacZ fusions are constructed to a comprehensive set of yeast genes. The fusions are generally constructed in a diploid cell of the a/a mating type to allow the introduction of dominant mutations by mating, though haploid strains also find use with particularly sensitive reporters for certain functions. The fusions are arrayed onto a grid separating distinct fusions into units having defined X-Y coordinates. The gene identification function 612 is performed by determining, for each reporter-tagged gene, a short sequence adjacent to the site of fusion. That sequence is then compared with the yeast genomic database to establish the identity of the gene. An index table 614 is established relating each gene in the matrix to the X-Y coordinate of the fusion construct for that gene.

The basal response determination function 616, is performed by measuring the basal response of each cell in the matrix under a variety of physical conditions, such as temperature and pH, medium, and osmolarity. This information is indexed against the matrix to form the reference response profile set 618 that will be used to determine the response of each reporter to any milieu in which a stimulus may be provided.

The compound treatment function 620 is performed by contacting each unit of the matrix with a test compound. Generally, a copy of the entire matrix is transferred to fresh medium containing the first compound of interest and the response is obtained for the entire matrix. In a reference subtraction function 622, the appropriate reference response profile is subtracted from the response profile, and the difference stored in the knowledge base as the first chemical response profile 624. Alternatively, the response profile is divided by the appropriate reference profile to yield an induction ratio. The process is repeated for compounds or mixtures of compounds 2 through N.

The gene mutation function 626 determines the response of the matrix to loss of function of each protein or gene or RNA in the cell introducing a dominant allele of a gene to each reporter cell, and determining the response of the reporter as a function of the mutation. For this purpose, dominant mutations are preferred but other types of mutations can be used. Dominant mutations are created by in vitro mutagenesis of cloned genes followed by screening in diploid cells for dominant mutant alleles. Alternatively, the reporter matrix may be developed in a strain deficient for the UPF gene functions, wherein the majority of nonsense mutations cause a dominant phenotype, allowing dominant mutations to be constructed for any gene. The data obtained identifies genetic response profiles 1-N 628. Note that N can be greater than the total number of genes in a species.

These data are subject to a sorting function 630 which sorts by individual gene response to determine the specificity of each gene to a particular stimulus. A weighting matrix 632 is established which weights the signals proportionally to the specificity of the corresponding reporters. The weighting matrix is revised dynamically, incorporating data from every screen into the N+1 weighting matrix. A gene regulation function 634 is then used to construct tables of regulation 634, 636 identifying which cells of the matrix respond to which mutation in an indexed gene, and which mutations affect which cells of the matrix.

The unknown stimulus matrix screen function 636, sequentially tests new chemicals or unknown compounds or unknown mixtures to identify output response profiles 642.

Figure 7 shows the steps performed in analyzing a gene reporter matrix output response profile for an unknown stimulus.

An output response profile 710, created for a new stimulus is shown subject to two alternative analysis paths. In the path shown on the left, elements 712-732, the new stimulus response profile 710 is subject to a comparison function 712 which compares the new stimulus response profiles with response profiles to known chemical stimuli. The comparison function 712 typically provides a comparison analysis in the form of an indexed report of the matches to the reference chemical response profiles or the genetic response profiles, ranked according to the weighted value of each matching reporter, the output response evaluation score (ORE SCORE). The comparison may proceed along a linear decision tree: The first query 714 is whether there is a match; if yes (i.e. perfect ORE SCORE), the output response profile identifies a stimulus with the same target 716 or targeted pathway as one of the known compounds upon which the response profile database is built. If no, the second query 718 is whether the output response profile is a subset of cells in the matrix stimulated by a known compound. If yes, then the new compound is a candidate 720 for a molecule with greater specificity that the reference compound. In particular, if the reporters responding uniquely to the reference chemical have a low weighted response value, the new compound is concluded to be of greater specificity. Alternatively, if the reporters responding uniquely to the reference compound have a high weighted response value, the new compound is concluded to be active downstream in the same pathway. If the output response profile is not a subset of cells in the matrix stimulated by a known compound, the third query 722 is whether the output overlaps the response profile of a known reference compound. If yes, the overlap is subject to a sort function 724 wherein it is evaluated quantitatively with the weighting matrix to yield common 726 and unique 728 reporters. The unique reporters are sorted 730 against the regulation tables 618, 620 and best matches used to deduce the candidate target 732. If the output response profile does not either overlap or match a chemical response profile, then the database is inadequate to infer function and the output response profile may be added to the reference chemical response profiles.

In the path shown on the right in Figure 7, elements 736-750, the output response profile of a new chemical stimulus is compared 736 to the genetic response profile for the target gene. The first query 738 is whether there is a match between the two response profiles. If yes, the deduction 740 is that the target gene is the presumptive target of the chemical. If no, the second query 742 is whether the chemical response profile is a subset of a genetic response profile. If yes, the deduction 744 is that the target of the drug is downstream of the mutant gene but in the same pathway. If no, the third query 746 is whether the output response profile includes as a subset a genetic response profile. If yes, the deduction 748 is that the target of the chemical is in the same pathway as the target gene but upstream. If no, the deduction 750 is that the chemical response profile is novel defines an orphan pathway.

The following examples are offered by way of illustration and not by way of limitation.

EXAMPLES

1. Transcriptional promoter-reporter gene matrix

A) Construction of a physical matrix stimulated with the drug mevinolin (lovastatin).

Mevinolin is a compound known to inhibit cholesterol biosynthesis. Initially, the maximal non-toxic (as measured by cell growth and viability) concentration of mevinolin on the reporter cells was determined by serial dilution to be 25 ug/ml. To produce a mevinolin-stimulated matrix, each well of 60 microtiter plates is filled with 100 ul culture medium containing 25 ug/ml mevinolin in a 2% ethanol solution. An aliquot of each member of the reporter matrix is added to each well allowing for a dilution of approximately 1:100. The cells are incubated in the medium until the turbidity of the average reporter increases by 20 fold. Each well is then quantified for turbidity as a measure of growth, and is treated with a lysis solution to allow measurement of β-galactosidase from each fusion.

B) Generation of an output signal matrix data structure.

Both the turbidity and the B-galactosidase are read on commercially available microtiter plate readers (e.g. BioRad) and the data captured as an ASCII file. From this file, the value of the individual cells in the reporter matrix to a 2% ethanol solution in the reference response profile is subtracted. The difference corresponds to the mevinolin response profile. This file is converted in the computer to a table indexed by the response of each cell to the inhibitor. For example, the genes encoding acetoacetyl-CoA thiolase and squalene synthase increase 10 fold, while SIR3, and LEU2, two unrelated genes, remain unchanged. The response of the reporter matrix to other compounds is similarly determined and stored as output response profiles.

C) Comparison of Output Signal Matrix data structure with an Output Signal Matrix database.

A physical matrix is constructed as describe above except the mevinolin is replaced with an unknown test compound. The resultant output response profile is compared to the response profiles of a library of known bioactive compounds and analyzed as described above. For example, if the test compound output profile shows both acetoacetyl-CoA thiolase and squalene synthase gene induced, then the output profile matches that expected of an inhibitor of cholesterol synthesis. If the output response profile has fewer other cells affected than the response profile to mevinolin, the unknown compound is a candidate for greater specificity. If the output response profile of the new chemical affects fewer other reporters than the response profile to mevinolin, and if the other reporters affected by mevinolin have a lower weighted value, then the compound is a candidate for greater specificity. If the output response profile has more different cells affected than the response profile to mevinolin, then the compound is a candidate for less specificity. In the case where mixtures of compounds are tested, the highest weighted responses are evaluated to determine whether they can be deconvoluted into the response profile of two different compounds, or of two different genetic response profiles.

2. Reporter transcript-oligonucleotide hybridization probe matrix: Construction of stimulated physical matrix and generation of an output signal matrix data structure.

Unlabeled oligonucleotide hybridization probes complementary to the mRNA transcript of each yeast gene are arrayed on a silicon substrate etched by standard techniques (e.g. Fodor et al. (1991) Science 252, 767). The probes are of length and sequence to ensure specificity for the corresponding yeast gene. typically about 24-240 nucleotides in length.

A confluent HeLa cell culture is treated with 15 ug/ml mevinolin in 2% ethanol for 4 hours while maintained in a humidified 5% CO_2 atmosphere at 37°C. Messenger RNA is extracted, reverse transcribed and fluorophore-labeled according to standard methods (Sambrook et al., Molecular Cloning, 3rd ed.). The resultant cDNA is hybridized to the array of probes, the array is washed free of unhybridized labeled cDNA, the hybridization signal at each unit of the array quantified using a confocal microscope scanner (Molecular Devices), and the resultant matrix response data stored in digital form.

3. Two-dimensional two-hybrid matrix

A) Construction of stimulated physical matrix.

The two-dimensional two-hybrid matrix is designed to screen for compounds that specifically affect the interaction of two proteins, e.g. the interaction of a human signal transducer and activator of transcription (STAT) with an interleukin receptor. Two hybrid fusions are generated by standard methods: each strain contains a portion of the targeted human STAT gene, fused to a portion of a yeast or bacterial gene encoding a DNA binding domain (e.g. GAL4:1-

147). The DNA sequence recognized by that DNA binding domain (e.g. UAS_G) is inserted in place of the enhancer sequence 5' to the selected reporter (e.g. lacZ). The strain also contains another fusion consisting of an intracellular portion of the targeted receptor gene whose protein product interacts with the STAT. This receptor gene is fused with a gene fragment encoding a transcriptional activation domain (e.g. GAL4:768-881).

B) Generation of output signal matrix data structure.

Both the turbidity and the galactosidase are read on commercial microtiter plate readers (BioRad) and the data captured as an ASCII file.

C) Comparison of output signal matrix data structure with database.

Data are analyzed for those compounds that block the interaction of the two human proteins by reducing the signal produced from the reporter in the various strains containing pairs of human proteins. The output is processed to identify compounds with a large impact on a reporter whose expression is dependent on a single pair of interacting human proteins. An inverted weighting matrix is used to evaluate these data as preferred compounds do not affect even the least specific reporters in the matrix.

All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference. Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.